RECOGNIZING EMOTIONS IN DIALOGUES WITH DISFLUENCIES AND NON-VERBAL VOCALISATIONS

Leimin Tian, Catherine Lai, Johanna D. Moore

School of Informatics, University of Edinburgh
s1219694@sms.ed.ac.uk, clai@inf.ed.ac.uk, j.moore@ed.ac.uk

ABSTRACT

We investigate the usefulness of DISfluencies and Non-verbal Vocalisations (DIS-NV) for recognizing human emotions in dialogues. The proposed features measure filled pauses, fillers, stutters, laughter, and breath in utterances. The predictiveness of DIS-NV features is compared with lexical features and state-of-the-art low-level acoustic features.

Our experimental results show that using DIS-NV features alone is not as predictive as using lexical or acoustic features. However, adding them to lexical or acoustic feature set yields improvement compared to using lexical or acoustic features alone. This indicates that disfluencies and non-verbal vocalisations provide useful information overlooked by the other two types of features for emotion recognition.

Keywords: emotion recognition, dialogue, disfluency, speech processing, HCI

1. INTRODUCTION

Emotions are vital in human cognitive processes. Emotion recognition has long been a focus in human-computer interaction research. State-of-the-art approaches for improving performance of emotion recognition often focus on identifying better feature representations. In this work, our goal is to identify knowledge-driven features that can improve recognition performance.

Psycholinguistic studies have shown that emotions can influence the neural mechanisms in the brain, and thus influence sensory processing and attention [9]. This in turn influences speech processing and production, which may result in disfluencies and non-verbal vocalisations. Therefore, we would like to investigate the usefulness of DISfluencies and Non-verbal Vocalisations (DIS-NV) for recognizing emotions in dialogues.

One of the most predictive feature sets identified for emotion recognition is the set of acoustic features based on low-level descriptors (LLD). However, in our previous work [7] on the AVEC2012 database [8] of spontaneous dialogues, DIS-NV features were more predictive than acoustic or lexical features for recognizing emotions. We would like to study whether our DIS-NV features remain predictive when the data contains both non-scripted and scripted dialogues. Therefore, we compare our DIS-NV features with LLD acoustic features and lexical features on the IEMOCAP database [1]. Our results show that although DIS-NV features are less predictive than acoustic or lexical features when used alone, they improve performance when combined with existing models.

2. METHOD

2.1. The IEMOCAP Database

The IEMOCAP database contains approximately 12 hours of audio-visual recordings from 5 mixed gender pairs of actors. Each conversation was about 5 minutes long. There are 10037 utterances in total, of which 4782 utterances were not scripted. When collecting the non-scripted dialogues, the actors were instructed to act out emotionally intense scenarios, e.g., telling a best friend that (s)he has been accepted into his/her most desired university.

Emotions were annotated at the utterance-level with a 1 to 5 integer score of the Arousal (activeness), Power (domination), and Valence (positive or negative) emotion dimensions. The mean score over all the annotations was used when the annotators disagreed with each other. We categorized the scores into three classes (<3, =3, >3) to have a clearer view of the relation between emotions and features, and to reduce the influence of imbalanced classes.

2.2. Features

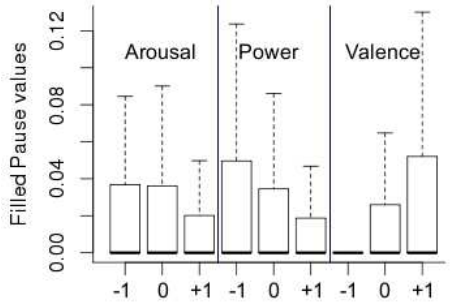
2.2.1. The DIS-NV Features

We studied 5 types of disfluencies and non-verbal vocalisations (DIS-NV): filled pauses (non-verbal insertions, e.g., “eh”), fillers (verbal insertions, e.g., “you know”), stutters, laughter, and breath. We choose them because they are the most common in the data, and they are relatively easy to extract from

transcripts. Disfluencies here refer to interruptions in the flow of speech production. Fluency of speech production may not always be the same with listener's perception of fluency [6]: Minor disfluencies may be ignored by the listener; In some cases, these tokens could also be perceived as part of a "fluent" utterance (e.g., using a filler at the beginning of an utterance while organizing sentences).

Feature values are calculated as the ratio between the sum duration of each type of DIS-NV and the total duration of the utterance, resulting in 5 DIS-NV features for each utterance. Descriptive statistics of filled pause features are shown in Figure 1 as an example. Utterances containing DIS-NVs are not very frequent in the IEMOCAP database (47.28% in the non-scripted utterances, 24.74% in the scripted utterances). To get a clearer view of value distributions, the statistics shown were computed on a subset of the data which contains all the utterances with disfluencies or non-verbal vocalisations (the DIS-NV subset).

Figure 1: Statistics of filled pause features.



2.2.2. *The Lexical Features*

The lexical features we extracted are 6 Point-wise Mutual Information (PMI) based features. PMI is a widely used measurement for the relation of words and emotions. It is based on the frequency of a word w having class label c, as shown: $PMI(c,w) = log_2(\frac{P(c|w)}{P(c)})$

To calculate PMI values, we first binarized all three emotion dimensions (<3, ≥ 3). PMI values of the scripted and non-scripted data are computed separately. The lexical features we proposed are calculated as the total PMI values of all the words in an utterance for each binarized emotion dimension, resulting in 6 lexical features for each utterance.

Example words with top PMI values are shown in Table 1. In the first column, "A-" represents unaroused, "A+" is excited, "P-" is dominated, "P+" is dominating, "V-" and "V+" represent negativeness and positiveness of emotion.

Table 1: Words with top PMI values.

	Non-scripted data
A-	Academy, Banking, Loan, Numb, Sleep
A+	Anger, Bloody, Flowers, Freak, Ruined
P-	Afraid, Beer, Error, Insane, Quit
P+	Bar, Chick, Duty, F*ck, Mad
V-	Abuse, B*tch, Die, Iraq, Unfair
V+	Australia, Cash, Dog, Snow, Tour
	Scripted data
A-	Lose, Non, Pets, Skip, Topic
A+	Bully, Cry, Gods, Jesus, Santa
P-	Bad, Cliff, Sacrifice, Sneak, Surprise
P+	Cry, Damn, Lose, Mad, Shut
V-	Ashamed, Crap, Hell, Sucker, Vile
V+	Delight, Eating, Gold, Loves, Wish

2.2.3. *The LLD Acoustic Features*

Our LLD acoustic features were the same as those used in the INTERSPEECH 2010 Paralinguistic Challenge extracted with OpenSMILE [3]. It represents a state-of-the-art feature set for emotion recognition. This feature set has been widely used as a reference for comparing emotion recognition feature sets and classification approaches.

There are 1582 LLD acoustic features, including those extracted by applying functionals (e.g., position of max) to low-level descriptors (e.g., MFCCs, F0, PCM loudness) and their corresponding delta coefficients, the number of pitch onsets, and the total duration of the utterance. Values are computed at the frame-level, with a window size of 60ms and a step of 10ms. Compared to DIS-NV and lexical features, LLD acoustic features overlook global characteristics of the utterance.

2.3. Experimental Settings

Our emotion recognition models were built with the LibSVM [2] classifier using WEKA [4]. We used the C-SVC approach with RBF kernel, and 10-fold cross validation. All features were normalized to [-1,1] before classification. Because of the imbalanced classes, we use weighted F-measure as the evaluation metric.

3. RESULTS AND DISCUSSION

The performance of different feature sets is shown in Table 2. "Mean" in the first row is the un-weighted average of the three emotion dimensions. In the first column, "DN" is the DIS-NV model, "PMI" is the lexical model, "LLD" is the LLD acoustic model.

Our results show that adding DIS-NV features to

Table 2: Performance on the full database.

Models	Arousal	Power	Valence	Mean
DN	0.363	0.407	0.328	0.366
PMI	0.483	0.483	0.332	**0.433**
PMI+DN	0.489	0.486	0.406	**0.460**
LLD	0.652	0.538	0.535	0.575

lexical feature set yields improvement on all emotion dimensions. This verified that DIS-NV features capture information neglected by the lexical content, thus helping with emotion recognition.

When used alone, DIS-NV features are less predictive than lexical or LLD acoustic features, which is different from our previous work. This may be caused by the different nature of the AVEC2012 and IEMOCAP database. Compared to the AVEC2012 database of spontaneous dialogues, disfluencies and non-verbal vocalisations are less frequent in the IEMOCAP database of acted data. To reduce such influence, we also performed experiments on the DIS-NV subset, as shown in Table 3.

Table 3: Performance on the DIS-NV subset.

Models	Arousal	Power	Valence	Mean
DN	0.470	0.453	0.329	0.417
PMI	0.500	0.467	0.316	**0.428**
PMI+DN	0.522	0.475	0.325	**0.441**
LLD	0.644	0.523	0.532	0.566
LLD+DN	0.645	0.525	0.533	0.568

Compared to using the full IEMOCAP database, when using this subset instead, performance of lexical features and LLD acoustic features has a small decrease, while performance of DIS-NV features increases greatly on all emotion dimensions. This verified the negative influence of infrequency of disfluencies and non-verbal vocalisations.

Adding DIS-NV features to lexical feature set remains helpful for all emotion dimensions. Adding DIS-NV features to LLD acoustic features only yields a small gain. The reason may be the great difference between the size of these two feature sets.

We further compared performance of individual DIS-NV features and LLD features with the CFS [5] method, which ranks features based on their individual predictiveness and their correlations with other features. DIS-NV features are always ranked among the top features, especially filled pauses, fillers, and laughter. This indicates that with a better fusion strategy, DIS-NV features may improve performance of LLD features greatly, by highlighting emotionally interesting segments.

Note that DIS-NV and lexical features describe data at the utterance-level, while LLD features describe data at the frame-level. In the future, with advanced fusion strategy that can combine feature sets at different levels with flexible weights, we may be able to combine information contained in these feature sets more efficiently and further boost performance of current emotion recognition models.

4. CONCLUSION

We proposed DIS-NV features measuring disfluencies and non-verbal vocalisations for recognizing emotions in dialogues. We compared their performance with lexical features and state-of-the-art LLD acoustic features. Our experiments on the IEMOCAP database show that using DIS-NV features alone is not enough for building a highly predictive emotion recognition model. However, these features contain information neglected by the lexical or LLD acoustic features. Thus, when fused properly, DIS-NV features may improve performance of current emotion recognition models greatly.

5. REFERENCES

[1] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42(4), 335–359.

[2] Chang, C.-C., Lin, C.-J. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3), 27.

[3] Eyben, F., Wöllmer, M., Schuller, B. 2010. OpenSMILE: the munich versatile and fast open-source audio feature extractor. *Proceedings of the international conference on Multimedia*. ACM 1459–1462.

[4] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter* 11(1), 10–18.

[5] Hall, M. A. 1998. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis University of Waikato Hamilton, New Zealand.

[6] Lickley, R. 2015. Fluency and disfluency. to appear.

[7] Moore, J., Tian, L., Lai, C. 2014. Word-level emotion recognition using high-level features. *Computational Linguistics and Intelligent Text Processing*. Springer 17–31.

[8] Schuller, B., Valster, M., Eyben, F., Cowie, R., Pantic, M. 2012. AVEC 2012: the continuous audio/visual emotion challenge. *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM 449–456.

[9] Vuilleumier, P. 2005. How brains beware: neural mechanisms of emotional attention. *Trends in cognitive sciences* 9(12), 585–594.